Nicholas Kong
February 23, 2009

Visualizing Statistical Analysis of News

CS260, Spring '09, Project Proposal

Background

Text is the most common form of visualization, and one of the most idiosyncratic. Since the advent of the Internet our access to text, and thus our difficulty to assimilate it all, has grown exponentially. This is especially true of news; we now have access to an unprecedented number of news sources, from the mainstream media to polemic bloggers.

While much work has been done in summarizing and condensing the news (e.g., NewsMap, Google News), to my knowledge no statistical analysis of large corpora of text has been used as the basis of a sensemaking visualization. In addition, the statistical analysis is itself novel. My research involves a collaboration with Professor Laurent El Ghaoui in EE and his StatNews group. The StatNews project is investigating statistical algorithms to analyze large text corpora, specifically news data. Their current approach involves Naive Bayesian analysis, whereupon the corpus is split into two sets (e.g., headlines that contain "Iraq" and headlines that do not), and each word is subsequently assigned a weight by the classifier. A positive weight indicates that a word is a significant predictor of Iraq in the headlines, whereas a negative weight indicates that a word is a significant predictor of the absence of Iraq in the headlines.

Approach

The main questions in this project are two-fold:

1. What does this analysis give us that co-occurrences do not, and
2. how do we display the results in such a fashion to reveal their significance?

In order to address either question, it is necessary to identify queries that our target users are interested in. As a starting point, we have started conversation with a social scientist about what she would be interested in using the tool for. She gave us queries such as

- Is Iran being portrayed as isolated?
- Is the word "spending" always used in the context of "social spending", or does it also encompass Department of Defense spending?

By discovering interesting queries from the target users, I will attempt to refine the existing visualization prototype (shown in Figure 1) and add additional features (such as the ability to compare two seed words, or the ability to dynamically alter the news source) in order to help answer those queries.

To address the first question I would also like to explore a similar visualization using just raw co-occurrences to discover if similar insights can be gleaned from both analyses.

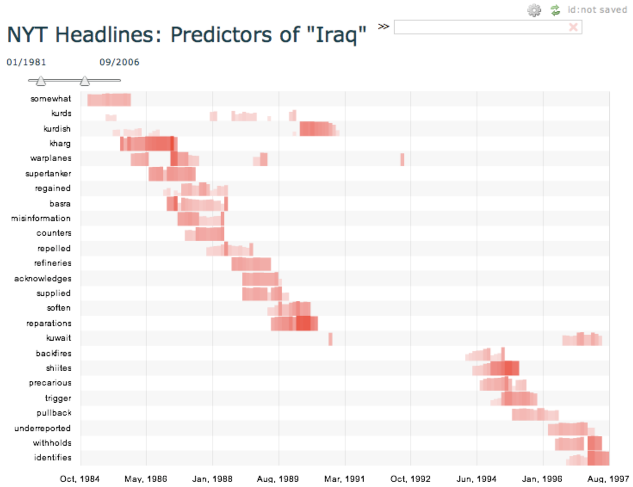

Figure 1: A view of the current visualization of an analysis of the New York Times headlines.

Themes from the course

This project principally relates to two course themes:

1. *Frames*
2. *Social Identity*

Frames

Instead of merely searching for coocurrences of, for example, "Iraq" and "evil", it will also be interesting to use this tool to compare different framings of the same or similar events. For example, we could compare the occurrences of "bailout" and "economic stimulus" to determine whether they are being used in mutually exclusive timeframes, or whether "bailout" is exclusively associated with "automobile" or "banking" whereas "stimulus" may be associated with "jobs" or "infrastructure".

Social identity

More tenuously, this project could potentially relate to social identity, specifically political social identity. By comparing news sources and left-/right- leaning blogs, the tool and analyses could be used to identify a prototypical "voice" for a certain social group. That is, we could find what words predict a right-wing blog versus what words predict a left-wing blog.

Project goals and assessment

Assessment will be primarily performed through case studies with social scientists or others who are interested in seriously analyzing this data. Since the design of the visualization will have been informed by one set of queries, we can determine the general efficacy of the visualization by presenting it to another group of users with another set of queries. Success will be mostly qualitatively determined: if users are able to find convincing evidence for their queries through the tool, we will have achieved our aims.